DOCUMENT RESUME

ED 220 507 TM 820 522

AUTHOR Arter, Judith A.
TITLE Out-of-Level Versus In-Level Testing: When Should We Recommend Each?
PUB DATE Mar 82
NOTE 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Equated Scores; *Evaluation Methods; Local Norms; Scores; *Testing Problems; Testing Programs; Test Interpretation; Test Norms; *Test Selection; Test Use; *Test Validity
IDENTIFIERS *Out Of Level Testing; Vertical Equating

ABSTRACT

Specific recommendations are made concerning the circumstances under which the benefits of out-of-level testing outweigh the problems associated with it. Topics explored are: various methods for deciding when a set of test scores is invalid and the utility of these methods for local evaluators, the accuracy of vertical scaling, and the usefulness of "chance level" scores as a criterion for out-of-level testing. The decision to test out-of-level is found to depend on the technical characteristics of the test that will be used, how a district will decide when out-of-level testing is required, the use to which the test scores will be put, and the practical requirements of testing in the district. A decision diagram for deciding when to test out-of-level is presented. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220507

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

OUT-OF-LEVEL VERSUS IN-LEVEL TESTING: WHEN SHOULD WE RECOMMEND EACH?

Judith A. Arter
Northwest Regional Educational Laboratory

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. A. Arter

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Introduction

Functional-Level Testing (FLT) is the process of assigning test levels to students on the basis of their skill level rather than on the basis of their grade level. Test publishers develop a test series with levels designed to match a range of skills at each grade or grades. These comprise the recommended levels for students at each grade. Although most students' functional levels are best served by giving the recommended levels, there are some students at each grade having a skill level not adequately covered by the in-level test. An increasing practice is to give these students a higher or lower level of the test--this practice is known as out-of-level (OOL) testing. FLT therefore allows testing at, below or above the publisher's recommended level.

For low achieving students, such as those in Title I or other remedial programs, FLT often means testing students OOL on a lower level of the test. This practice is appealing for several reasons (Farr, 1969; Raffeld, 1980; Roberts, 1976, 1980). When a test is too hard, a student may become frustrated and stop taking the test or may simply guess on most items. Such scores would not necessarily represent what a student really knows and are therefore of questionable value. Also, if the content of a test does not match what the student is likely to be taught, test scores are not useful for planning instruction. Testing OOL can

TM 820 522

Paper presented at the annual meeting of the American Educational Research Association, March 19-23, 1982, New York, New York.

often produce a better content match to curriculum. Finally, tests that are too difficult can produce misleading program evaluation results by placing an artificial floor on test scores (Fry, 1971; Wick & Ward, 1977; Haenn, 1978).

Although appealing for the above reasons, FLT can run into problems. Leaving aside some practical concerns such as how to schedule testing when more than one test level is to be given in the same classroom at the same time, interpretation problems and anomalous results can occur when testing OOL, especially if levels are changed between pretest and posttest. For example, one longitudinal study (Ozenne, 1978) showed that growth curves over time were smooth except for the year when test levels were changed. In this study, the fourth grade norm tables for two adjacent levels of the test showed one set of norms to be "easier" than the other for the group examined. Another example is one test series on which percentiles obtained using actual OOL conversion tables were different from those obtained when the test's vertical scaled scores were converted to percentiles using the in-level tables (Stewart & Tallmadge, 1980).

Such anomalous results occur because, for the group tested, the scores used to link the levels of the test are not comparable between levels. In order to explain this it is necessary to note that test publishers link levels of their test series through various vertical scaling or vertical equating techniques which results in a vertical scaled score (sometimes called expanded scaled scores, scaled scores, growth scores, etc.). These scores are supposedly comparable between levels--that is, any group taking two adjacent levels of the test should obtain the same average vertical scaled score (VSS). Comparability of

scores between test levels is of concern in out-of-level testing because scores on the out-of-level test must be converted to in-level derived scores if it is desired to obtain percentiles or combine scores among students taking different test levels. Comparability of scores between levels is also necessary for longitudinal testing--testing in which it is desired to compare progress over several years. The adequacy of the VSS (that is, the comparability of VSSs betweeen levels) can be affected by at least three factors: the vertical scaling procedures used by test publishers to equate the levels of a test series, unreliability in certain portions of the score distribution, notably in the extremes, and the similarity of the local population to the norm group. The latter two are beyond the control of the test publisher.

Thus, it appears that error in measuring the skill level or relative standing of students can occur whether one tests in-level or out-of-level. Error from in-level testing can stem from a test that is too hard or does not have the proper content, while error in out-of-level testing stems from problems in the vertical equating, unreliability of scores in the extremes of the distribution and characteristics of the local population. The decision as to whether to test out-of-level or not really depends, then, on an estimation of which of these sources of error is likely to be <u>least</u> problematic in any given situation.

This paper reviews the literature in order to make recommendations concerning the situations in which the benefits obtained from OOL testing outweigh any added sources of invalidity stemming from OOL testing. An informed decision on whether to test OOL depends on whether a set of test scores from a given level is accurate, whether the content of any given level match what is taught, the technical adequacy of the VSS of the test

to be used, how the district plans to use the test scores obtained, and various practical constraints on testing in the local situation. Although all of these topics are of concern and will be touched upon in this paper, most of the review will concentrate on the first and third items--whether a set of scores is accurate, and the technical adequacy of the VSS of various tests. Matching test content to curriculum and practical concerns when testing out-of-level are covered adequately elsewhere (see, for example, NWREL TAC materials on Test Selection and Functional Level Testing, respectively).

With respect to judging whether a set of test scores is adequate, this paper reviews 10 methods for deciding when individual or group scores are valid measures of achievement. The usefulness of the various methods for local users is discussed, and a procedure is recommended for judging test score validity. Some time is spent on "chance level scoring" because of the emphasis placed on chance level both as a criterion for deciding when to test OOL and as a criterion for judging the adequacy of VSSs. The second section of the paper reviews studies which examine the adequacy of VSSs. The conclusion syntheses the results from all of these perspectives plus content, use and practicalities into a set of guidelines for deciding on when to test OOL.

Determining When a Set of Test Scores is Valid

One of the major reasons for considering OOL testing is that it is felt that an in-level test is too difficult. Scores on such a test might be more an indication of random guessing behavior than true measures of achievement. A determination that a test level is too difficult and results in nonmeaningful scores would certainly tend to balance the error

equation in favor of OOL testing. The problem then becomes how to tell whether a test is too difficult, or alternatively, whether a set of test scores is a valid representation of achievement.

The emphasis in this section is to explore the various methods that have been suggested for determining whether test scores are accurate representations of achievement. If a set of test scores does turn out to be invalid, and the cause appears to be a test that is too difficult, then many of the same methods used to determine validity can also be used to decide on the proper level of the test. Some of the methods emphasize making judgments about individual scores while others emphasize judgments about the average score for a group of students. Each method will be so identified.

"Chance"

The most wide-spread criterion for judging when a test score is invalid is the "chance" level score. This is a score that could be generated by random guessing on a test--the so-called Orangutang Score (Fry, 1971). Such a score obviously does not tell much about a student's skill level.

The chance level on a multiple choice test is often defined as the total possible number of items divided by the number of choices per item. Most multiple choice tests have four choices per item. Therefore, a chance level score is 25% of the total number possible. The chance level is sometimes amended to include a confidence interval, amounting to a cutting score which would not happen by chance guessing a given percentage of the time.

The exact probability of a score or group average occuring by chance depends on the number of students taking the test, the number of items in the test, and the spread of scores. Since determining the exact probability that a score occurred by chance guessing can be lengthy, especially for practitioners unfamiliar with statistics, several authors have proposed "rules of thumb" for determining when the chance level has occurred (Barnes, 1977; Ayrer & McNamara, 1973; Farr, 1969; Roberts, 1976 & 1980; Fry, 1971). In general, these authors recommend that one must obtain around 30% to 75% of the items correct in order to have the score be considered valid.

There have been several objections to using chance level as a criteria for judging whether test scores are valid. First, the rules of thumb hold only when students either legitimately try on all items or randomly guess on all items (Stewart, 1980). What may actually happen is that a student will try on some items and guess on others, may rule out one or more choices per item leaving only two or three instead of four to guess between, or not complete the test. The first situation would mean that only part of the test score is invalid--but, which part? The second situation might mean that a student would get around 50% of the items correct and still be guessing. The third situation might yield a test score that is in the chance range and still be valid. In addition, many authors believe that people do not typically guess on tests (Donlon, 1971; Durost, 1971; Ebel, 1968; MacGinitie, 1971; MacRae & Green, 1971).

There is also considerable evidence to show that scores in the chance range tend to be valid indicators of achievement. There have been several tactics used to explore this issue.

Easy and Hard Tests. The most frequent tactic has involved examining the test scores of a group of slow students on two tests (Angoff, 1968; Boldt, 1968; Cliff, 1958; Levine & Lord, 1959). Although both tests areassumed to measure the same skill continuum, one test is usually easier than the other. The hard test is used to generate a sample of students scoring at or below chance. The easy test is designed to minimize chance level scorers and hence be a true measure of the relative abilities of students. The scores of one test are then regressed on the scores of the other. The scores on the easy test for those students who scored at or below chance on the hard test are examined to see whether different scores at or below chance correspond to separate scores on the easy test. If so, chance scores can be thought of as having predictive validity--a student who gets a score higher in the chance range is truly more able than a student who gets a score lower in the chance range. In four of the five groups studied in the four papers listed above, chance level scores were predictive. However, regression lines in the chance range were generally not colinear with those in the nonchance range and the variance of prediction was generally larger in the chance range. Thus, it appears that although group averages at or below chance can be thought of as valid measures of relative status in students, there is an increase in the amount of guessing in such scores and they may not be accurate measures of actual status.

R per bis. A second tactic used to examine the validity of scores in the chance is the "r personal biserial" ($r_{per\ bis}$) (Donlon & Fischer, 1968; Donlon & Rindler, 1979; Fischer, 1970). In the studies using this measure to examine validity of chance level scores, it was found that as test scores decrease $r_{per\ bis}$ also decreases slightly, showing a

greater tendency to guess. However, most students still had a $r_{per\ bis}$ that was significantly greater than zero. These results are consistent with those using hard and easy tests.

Distributions. A third tactic compares the distribution of chance level scores to the distribution that would be expected by random guessing (MacRae & Green, 1971; Stewart, 1981). MacRae & Green used the 200,000 students in the norming sample for the CAT (1970). Stewart used the norming sample of the 1978 MAT. For each study, the two distributions were significantly different at all test levels examined. The authors concluded that the low scores did not fit a random behavior pattern.

Other Evidence. Stewart (1981), as discussed previously, used several methods to identify potentially invalid group and individual scores. Some of the results of comparing these methods bear also on the issue of the validity of chance level scores. Stewart's results generally showed that (1) pretest-posttest correlations for students scoring below chance on the pretest were generally lower than those for the entire group, even after correcting for restriction in range, (2) although 2/3 of the students showing large pretest-posttest gains had a pretest NCE of less than 45, large pretest-to-posttest gains were also recorded by many students who scored toward the center of the pretest score distribution, (3) no more than about 1/3 of the pretest chance-level scorers showed gains or losses of 20 NCEs or greater, and (4) although students with lower raw scores tend to omit more items, some students who score near the middle of the pretest raw score distribution also do not respond to significant numbers of test items.

Hills and Gladney (1968) correlated high school chance range SAT scores to GPA after one year of college and regressed GPA on SAT scores. They found that regression lines were not significantly different for students above and below chance on the SAT.

Conclusions. The studies on the validity of scores in the chance range seem to indicate that students, even low achieving students, do not invariably display true random behavior on tests. Guessing increases as test scores decrease, but differences in group averages in the chance range appear to be indicative of actual differences between groups of students. However, these studies only show predictive validity of chance level scores as relative standing for groups, they do not demonstrate that they are accurate measures of actual status. In addition, the fact that group scores are predictive does not imply that a score for any individual student is predictive. These results would indicate that chance level could be used as a flag for the possibility of an inaccurate measure of ability, especially for individual students. But chance level should not be taken as proof of an inaccurate score (Wilson & Donlon, 1980; Stewart, 1980), should not automatically be used as a criterion for deciding when to test OOL, and should not uncritically be used as a criterion to judge the success of OOL testing with low students.

Patterns of Responses

The presence of identifiable patterns in a student's responses to ultiple choice questions might be a good indicator of an invalid test score for individual students (Schmidt, Liberty & Butera, 1979; Stammon, 1979; Stewart, 1981). For example, a student might mark the same choice all the way through, or draw a picture by filling in choices. Sometimes patterns are harder to detect. For example, a student might use the

pattern a-b-d-a-c-d-b. Is this a pattern or not? Schmidt, et. al. (1979) used the criterion that a total test score is invalid if half of the student's correct responses fell in such patterns. In one study, they found that only 3% to 5% of the test scores are invalid using this method. This was a smaller number than any of the other methods they investigated. Stewart (1981) likewise found that this method identifies the smallest number of scores as invalid. The small number of scores identified as being invalid is probably due to both the difficulty of identifying patterns, and the fact that humans may not make up patterns when they guess.

Chi-square

This method works best when the items on a test are easily divided into hard and easy items. On a spirally designed test (where test items at the end are more difficult than test items at the beginning), the first half are considered easy. One could also look at p values to divide items into those that are easy and hard. A two-by-two table is set up for an individual as shown below (Stammon, 1979; Donlon, 1971):

	hard	easy	
correct	5	18	23
incorrect	20	7	27
	25	25	50

A chi-square statistic is then calculated for this matrix. If the chi-square turns out not to be significant then the score is to be considered invalid. This is because one would expect that a student would get more easy items right and more hard items wrong if they were making a valid attempt on the test (as in the table above). A nonsignificant chi-square indicates that the student missed just as many

easy items as hard items, which is a good approximation of guessing behavior.

Personal Biserial

The personal biserial index, $r_{per\ bis}$ (Fischer, 1970; Fischer, 1961), is the correlation between a person's item scores (zero or one) and p for each item (the proportion of people in the norming sample that got the item right). This method is very similar to the chi-square method in that if a test score is valid one would expect that more easy items would be correct and fewer hard items would be correct. This pattern would result in a high, positive, $r_{per\ bis}$. As guessing behavior increases, the correlation goes toward zero.

Regression

Stammon (1979) devised a regression equation to predict when a test score can be considered invalid. He first defined an invalid test score by the chi-square method. Since this appeared to be a time-consuming method, he searched for other variables that would predict the results of the chi-square method. The equation he ended up with is:

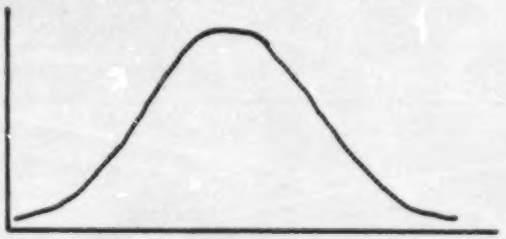
$$Y = .46 + .1X_1 + 1.43X_2$$

where X_1 is the percent correct of the number attempted by a student and X_2 is the number of correct responses on the second five items. To use the equation, one would plug in a students' percent correct and score on the second five items. If the result is greater than .5 the test score can be considered to be valid; if the result is less than .5, the test score is to be considered invalid.

Schmidt, et. al., (1979) found that about 40% to 50% of the test scores they examined were identified as being invalid using this method.

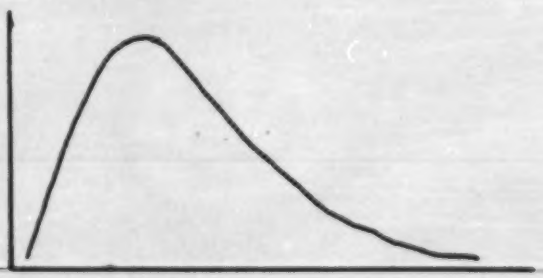
Shape of the Score Distribution

Theoretically, if we assume that the distribution of a trait being measured is normal, and the sample of students tested is representative of the population for whom the trait is thought to be distributed normally, one should obtain a bell-shaped (normal) curve when one looks at the frequency distribution of scores for a group.



Thus, a few persons would get very low and very high scores and lots of persons would get scores in the middle range.

One indication of inaccurate test scores might be nonnormal distributions of scores (Roberts, 1976; MacRae & Green, 1971; Doherty, 1976; Haenn, 1978). Test scores would pile up toward the bottom if the test was too hard, if a lot of students were guessing, if the time allowed for the test was shorter than that specified in the manual, if distractions occurred, etc. Test scores would pile up toward the top of the test if the tests were too easy, if the answer sheet circulated prior to testing, etc. A skewed distribution is depicted below.



One could actually plot the number of students which got each test score to see whether the distribution looks skewed. Or one could compare

the mean score in a group to the median score. For a normal distribution of scores the mean and the median are the same. When test scores pile up at the bottom the mean can be substantially larger than the median. When test scores pile up at the top, the mean is substantially smaller than the median. Roberts (1980) suggests that if the mean is more than 1/3 standard deviation greater than the median, there is a floor effect, and if the mean is more than 1/3 standard deviation less than the median, there is a ceiling effect.

Correlation Test

This method compares the pretest-posttest correlation of a subgroup to that for the entire group (Stewart, 1981). The correlation for the subgroup is corrected for restriction in range. For example, the subgroup could be Title I students and the entire group could be the rest of the school population. Or, the subgroup could be Title I students scoring below chance and the entire group could be all Title I students. If the correlation for the subgroup is smaller than that for the entire group, then additional error variance may be present in the subgroup scores.

The author found that, using the MAT norming data, the correlation for groups below chance did not approximate that for the entire groups. However, problems appear when using this method for small subgroups--corrections for range restriction often change the entire meaning of the correlation. For example, Stewart (1981) reported changes such as -.20 to .58 and .16 to .65 with Ns over 100, and -.71 to 1.00 when N is four. Such shifts are difficult to interpret.

Gain Scores

The gain score test (Stewart, 1981) is a method for identifying individual student scores which may be invalid. As a criterion for

deciding when a gain score was too large, Stewart used a gain score that fell more than three standard deviations away from the mean gain. If the gain is this large it is likely that the pretest and or the posttest is invalid. Using those students in the MAT norming sample who had both fall and spring scores, the author identified between six and 11 percent of the scores as potentially invalid.

Initial Items

Since the first five items on an achievement test are usually the easiest, Stewart (1981) felt that any student not answering or missing three or more items in the first five might indicate an inaccurate score. Unfortunately, only seven students in the MAT norming sample failed to omit three of the first five items. This is therefore, not a very discriminating method for identifying invalid scores.

Frequency of Omitted Responses

Stewart (1981) looked for large numbers of omitted responses in individual test scores. He found that many students omitted items. As expected, there were fewer average omitted items as raw scores increased. Also the number of omitted items increased sharply at a raw score of 22% correct--the mean number of omits for each score below 22% correct ranged from 17 to 50, and for scores between 22% and 29% correct the mean number of omits ranged from 7 to 11.5. The mean omits for raw scores greater than 29% correct ranged from 0 to 5.5.

Conclusions

In general, the methods for identifying when test scores may be invalid vary as widely as the resulting scores they identify (Stewart, 1981; Schmidt et. al., 1979). Some methods for identifying when scores may be inaccurate (chi-square, $r_{per\ bis}$, regression, patterns of

responses, correlations test, and shape of the score distribution), are harder to apply than others (chance level rules of thumb, initial items, number of omits, gain scores). This puts some of the more promising tests (especially $r_{per\ bis}$) out of range for many users.

Another concern is the wide variation in the number of scores identified as being potentially invalid by the various methods. There is no outside criterion by which to judge which method is most accurate. $R_{per\ bis}$ appears to have promise. Since this takes time to apply, a workable solution might be to use the easier methods as initial screening flags with the harder methods used to refine judgments. Some methods, however, probably should not be used at all—initial items, correlations when small Ns are involved, and patterns of responses. These methods appear to be nondiscriminating and/or misleading.

Adequacy of Vertical Equating

Vertical scaling is the process of equating test levels so that scores on one level can be compared to scores on another level. As discussed in the introduction, this comparability is of concern in out-of-level and longitudinal testing because scores on one level of the test must be converted to scores on the other level if it is desired to obtain percentiles, combine scores, or compare scores among students taking different levels. To the extent that equating is inadequate, one would argue more strongly for in-level testing.

One problem is that there is no universally agreed upon definition of what an adequate VSS should look like. Studies which attempt to look at the adequacy of vertical equating generally proceed by establishing some logical criterion for what would be expected if equating were adequate, and then attempting to determine whether this occurs in practice. We

have identified two categories of criteria--characteristics that would be expected of the norms tables themselves, and how derived scores for a group of students taking more than one level of the test should interrelate.

Characteristics of the Norms Tables Themselves

There were two studies which attempted to make predictions concerning how the norms tables should function if vertical scaling is adequate.

Stewart & Tallmadge (1980) postulated that when a test series has both out-of-level norms tables and a VSS, the derived scores obtained by using either should be the same. They plotted VSS against percentile for adjacent levels of the CAT and CTBS using publisher's or Sustaining Effects Study norming samples, respectively. They found that the CTBS has fairly large discrepancies between the scores derived by each method while the CAT showed a very close correspondence. Discrepancies such as on the CTBS would have severe implications for interpretation of OOL test scores. Users should check their norms tables.

Butera and Raffeld (1979) felt that in order to maintain the logic of OOL testing, students should have to score above the chance level on the OOL test in order to obtain an in-level percentile above one, and thatstudents should not be able to score above the 40th in-level percentile when taking an OOL test. When they examined the SRA, CTBS, CAT, SAT & Gates, they found that in 49% of the test levels considered, students could score above the first percentile and still score below chance (30% correct) on the OOL test. They also found that by scoring 75% or more on the OOL test, a large portion of the tests showed an in-level percentile above 40. The degree to which this is taken to be critical of vertical scaling depends on the validity of chance scoring

and the extent to which one accepts the authors' other criteria for how tests should function.

Comparability

The purpose of VSSs is to provide a way of telling what score on one level of a test would correspond to which score on another level. That is, if a person took one level of the test and obtained a certain score, what score would he have been likely to obtain had he taken another level instead? This purpose of VSSs leads some authors to the notion of "comparability" as a criterion to judge their adequacy. Although, the single most popular criterion for vertical scaling adequacy was comparability of test score or test score averages between levels, the details of exactly what comparability entails differed somewhat between authors.

Equal Averages. One type of comparability study examined the test scores obtained by a group of students on two levels of a test series. The criterion for equating adequacy and definition of "comparability" was that average scores on the two levels should be the same. Seven authors examined equivalency of standard scores or percentiles (Crowder, 1978; Long, 1979; Ozenne, 1978; Pelavin & Barker, 1976; Powers, 1978; Slaughter & Gallas, 1978; Stewart & Tallmadge, 1980) and four authors consideredequivalency of grade equivalents (Barnes, 1977; Fisher, 1961; Long, Schaffram & Kellog, 1977; Young, 1972). Results of these studies are not consistent--some show that score averages between levels are the same, some show that the higher level yields higher means, some show that the lower level yields higher means, and some show mixed results depending on the grade level of students, the test and the subtest used.

These results have frequently been taken to mean that comparability of VSSs cannot be assumed, and therefore results between levels cannot be compared. These results have lead to various recommendations-for example, one recommendation is to always test in-level, and another is to test at functional level but do not change levels between pre and posttesting.

Other individuals have found logical flaws with the conclusion of VSS inadequacy because other factors besides VSS inadequacy could produce these results. One such major factor is whether the authors controlled for the ability of the students. Some authors administered two test levels, both of which were within the functional range of the students, while others administered an in-level test and another test designed to be at the students' functional level. The problem with the second procedure is that measures on a harder level may be less accurate with respect to actual standing because of guessing or other factors. Therefore, one would not necessarily expect scores from the two levels to be the same (Raffeld, 1979; Temp, 1979). (This argument is not altered by the studies considered previously in which chance level scores were shown to be predictive. Even if chance scores on a harder test do correspond to separate scores on the easier test it does notimply that these two sets of scores are the same, only that they rank students in the same manner.)

Table 1 shows a breakdown of studies by those that considered difficulty and those that did not. Studies that considered test difficulty (Slinde & Linn, 1977; Crowder, 1978; Ozenne, 1978;) also appear to show no consistent pattern of results. This means that when students were given two test levels, both of which were within their functional range, derived scores were still not consistently the same.

Table 1
Summary of FLT Comparability Studies

Both Levels
Functional

Study	Grades/N*	Test(s)*	Score	Range	Results**	Problems
Equal Averages						
Barnes, 1977	1-12/N.S.	ITBS	G.E.	-	Same	
Crowder, 1978	3,5/300	SAT/MAT	SS	+	Mixed	
Fisher, 1961	3-6/6940	ITBS	G.E.	-	Same	Tested OOL and in-level two months apart and linearly adjusted for difference
Long, 1979	2,3,4/482	Gates/Botel	ESS	-	Different: OOL derived scores greater than in-level	
Long et. al., 1977	2,3,4/482	Gates/Botel	G.E.	-	Mixed	
Ozenne, 1978	4/N.S.	CAT	NCE	+	Different: OOL derived scores greater than in-level	
Pelavin & Barker, 1976	4,5,6/12 classes	MAT	SS	-	Mixed	Order effect: OOL given first
Powers, 1978	4,7,9/89	CTBS	ESS	-	Mixed	
Slaughter & Gallas, 1978	6/101	CAT	NCE	-	Different: OOL derived scores lower	

Study	Grades/N*	Test(s)*	Score	Range	Results**	Problems
Stewart & Tallmadge, 1980	2-6/72,500	CAT/CTBS	SS/NCE	-	CTBS: Same CAT: Mixed	CAT: different students took the two levels
Young, 1972	3,5,7/3900	MAT	G.E.	-	Mixed	
District Averages						
Ayrer & McNamara, 1973	4-8/7500	ITBS	G.E.	-	Mean G.E. tended to decrease as number of students tested OOL increased	Different students in each grade each year
Barnes, 1977	1-12/N.S.	ITBS	G.E.	-	Mean G.E. decreased during 1st year of OOL testing	Different students in each grade each year
Wick, 1977	1-12/280,000	ITBS	G.E.	-	Mean G.E. decreased during 1st year of OOL testing	Different students in each grade each year
Other						
Applebaum, 1979	N.S./6,000	ITBS	G.E.	N.A.	Same	
Divgi, 1980	N.S./2,000	N.S.	G.E.	N.A.	Different: Higher level derived scores greater than lower level	
McCormick, 1979	3/668 schools	Various	N.S.	N.A.	In-level test inflates scores of low students	
Slinde & Linn, 1977	4,5,6/N.S.	Various	G.E.	+	Mixed	

*N.S. means not specified

**Same: OOL derived scores about the same as in-level derived scores.
Mixed: In same grades OOL and in-level derived scores are the same, and/or in other grades they are different.

Another factor that can affect results is test content. Raffeld (1980) showed that with several levels of the CAT, there are different skill emphases in reading comprehension and vocabulary. Jenkins & Pany (1979), Armbruster, Stevens & Rosenshine (1977), and Porter, Schmidt, Floden & Freeman (1978) demonstrated that for a given curriculum various tests could yield different measures of achievement. The relative match between a curriculum and various test levels may have an effect, therefore, on the comparability of averages--one level may measure changes due to the curriculum better than another. Not enough information is given about curriculum in the studies in Table 1 to determine whether this could have affected their results.

The particular test used in a study and differences between the local group of students used in the studies and the norming populations of the test could also be confounding factors when examining comparability of scores. There appeared to be no consistent differences between tests, although it was difficult to make generalizations because six different tests were used in the eleven studies. Not enough information was provided about students to determine whether this made a difference.

District Averages Between Years. A second type of comparability study examined school-wide scoring patterns between those years in which everyone was tested in-level and those years in which FLT was instituted (Ayrer & McNamara, 1973; Barnes, 1977; Wick, 1977). The criteria depended on the author. Some felt that mean grade equivalent scores should be the same with different levels while others thought that mean grade equivalents should decrease when FLT is instituted because the floor is removed. Generally, these studies found that means decreased during the first year of FLT and that means decreased with an increase in

the number of students being tested OOL. (It is interesting to note that evidence of decreasing grade equivalents has been taken as both supportive of and a blow to the issues of whether results are comparable. On the one hand, if OOL and in-level tests yield different GEs then scores are not comparable. On the other hand, in most studies FLT was begun because it was seen that in-level tests were too hard and a floor effect resulted. Therefore, the decrese in GEs would seem a natural byproduct of the removal of the floor).

These comparability studies are also affected both by test level appropriateness and test content. Additional internal validity problems involve the facts that testing either occurred a year apart or that different sets of students were being compared. The former problem was that the in-level test was given one year as part of the district's regular testing program and the OOL test was given the next year. Therefore, differences in grade equivalents may be due to other events in the school than changing test levels. The latter problem is that averages in a single grade were compared between years. Thus, different sets of students were involved.

Anchor Test. Slinde & Linn (1977) used anchor test data to assess the comparability of grade equivalents between levels. The authors used data from the Anchor Test Study (Bianchini & Lovet, 1974) in which one single test was "anchored" to several others in order to equate them. The anchor test had levels which covered more than one grade each, while some of the other tests had separate levels for each grade. This allowed the authors to convert raw scores on the broad levels of the anchor test to raw (and thence to derived scores) on each of two adjacent levels of a second test with narrow levels. The authors found inconsistent

relationships between derived scores--sometimes the higher level showed a greater derived score and sometimes the lower level showed a greater derived score. Many differences were small.

RASCH. Applebaum (1979) and Divgi (1980) used Rasch equating to assess comparability of scores between levels. In each study, the author placed the raw scores of students taking different levels of a test onto a common scale. This was possible because there was an overlap of several items between levels. The author then converted the equivalent raw score on two test levels to derived scores as given by the original norms book. Both authors found that the match was better in the middle than at the extremes of the score range.

Regression. One final approach to assessing comparability was undertaken by McCormick (1979). He divided 437 schools taking in-level pre and posttest into two groups--those with average scores above 30% correct on the pretest and those with average scores at or below this chance cutoff. He then regressed post on pretest scores for the above-chance schools and used the resulting equation to predict the "true" pretest score for the schools below chance by plugging in their posttest averages. He found that in the chance schools the pretest was inflated by about five percentile points. The author interpreted these results to mean that pretest averages were overestimated and thence gains for these projects were underestimated. Hence in-level scores were not comparable to what students would have obtained if they had been tested at their functional level.

Summary and Synthesis

Studies which attempted to look at the practical adequacy of VSSs tended to examine characteristics of the norms tables that would be

expected if VSSs were accurate, and some notion of how the tests would function in use if VSSs were accurate. The primary characteristic of the norms tables expected is if both OOL norms tables and a VSS are present in the same test series, the derived VSSs obtained should be the same regardless of the vertical equating method. Stewart (1981) showed that this was not necessarily true for all tests, especially in the "unreliable range" of the test--the extremes of the score distribution.

Studies which looked at how tests would function in practice if VSSs were accurate tended to examine whether groups of students would get the same derived scores if they were given different levels of the test. Empirically, there were many instances where the groups examined did not obtain close derived scores between the levels or did not exhibit the pattern of differences between derived scores that was expected. Although this was empirically the case, it is not devastating to the theoretical concept of a VSS since many other factors besides an inadequate VSS can contribute to these findings. Two of these factors are whether the content of one level of the test series happens to match the particular local curriculum of the students examined better than the other level, and the extent to which the local group has the same learning characteristics of the norming group. Even if the equating is adequate in the norming population, differences in content and student characteristics may make the equating less vaid for a particular local population.

Another factor complicating the result was that many of the studies used groups for which one level of the test was within their "functional" range of ability while the other test was in-level. One would not necessarily expect that the scores from the two levels would be the

same. In other words, if the scores from the two tests did turn out to be the same, why bother to test OOL? The reason for testing OOL to begin with is because one did not believe that the in-level test was appropriate. In fact, there is an indication in the studies reviewed that there was a greater chance of obtaining smaller differences in derived scores in the middle range of the distributions of both levels, while more disparate and inconsistent relationships were noted when groups tended to be in the extremes of one or both of the distributions. This again tends to support the conclusion that the VSS, in general, tends to function best within the overlapping reliable range of both levels.

But where does this leave us? Most of the students for which we would consider OOL testing fall within the extremes of the score distribution (especially at the low end) in the area where measurement tends to be less reliable. In this range it has been empirically observed that derived scores from the two levels are not the same. Now the question becomes which score to trust?

There exists no outside standard or criterion for deciding which derived score is the more accurate. Therefore, any arguments as to which level should be used based only on observed evidence as to VSS adequacy can lead to very confusing self-referent results. To demonstrate the complexity of the decision a short dialog is presented below between a proponent of in-level testing (IL) and a proponent of OOL testing.

IL: We should only test in-level because studies on VSS accuracy with low students show that one does not obtain the same result with both levels, and anyway, low scores have been shown to be predictive.

OOL: Low scores may be predictive but that does not imply that they are accurate measures of status--the studies only show that students are rank-ordered properly. Anyway, studies show that test scores are less reliable in the low range probably due to guessing. Since the tests are too hard for these low students, we should always test OOL.

IL: If you test OOL because the in-level test is too hard, then how can you convert back to an in-level derived score and say that it is more reflective of actual achievement? That is, if we accept your argument, then an accurate OOL score is to be converted to unrealiable in-level scores using a VSS that is inaccurate in the low range of the in-level test. Given this, testing in-level is at least less unambiguous.

OOL: But you misrepresent my argument. If low scores have been shown to be predictive then it might be the case that equating is more or less accurate in the norms group, even at low scores. If we then test OOL with a local group because their scores have been shown to be invalid, then the more accurate OOL scores would be converted back to in-level derived scores using a good VSS table.

IL: Does this make sense? If the in-level scores were invalid for a local group, then how can you make sense of what the in-level equivalent of an accurate OOL score is? That is, they would not have gotten that score had they taken the in-level test.

Self-referent arguments are very difficult to resolve because a proponent's original assumption about the correct course of action implies assumptions about the results which can then be taken to be proof of the original assumption. Since circularity is certainly the case in

this situation, we maintain that it is inherently impossible to resolve the question of whether in-level scores derived from the in-level or OOL test are more accurate in an abstract sense for students scoring in the unreliable range of a test. This does not mean that adequate vertical equating should be ignored when deciding on test levels, only that if a certain amount of care has been devoted to developing the VSS, by the test publisher, then known problems with equivalences can be weighed only in light of other needed criteria for a test such as test content and test score use.

Discussion

The question posed at the beginning of this paper was whether we should recommend OOL testing. The paper attempts to point out that there is no simple answer to this question. The decision to test OOL really depends on whether in any given situation the increase in validity and usability of the test scores obtained by testing OOL offsets any new validity problems which arise because of testing OOL. Two factors which affect this tradeoff were discussed in detail in the paper--the extent to which a set of test scores is a valid measure of achievement, and the extent to which the vertical equating of tests is adequate. Other considerations mentioned briefly were the use to which the test scores would be put, the content of the various test levels, and whether OOL testing is a practical alternative in a district.

As a way of summarizing and interpreting the concerns presented in this paper, we propose the following decision diagram for deciding when to test OOL. In the diagram each of the major decision boxes is numbered so that it can be referenced in the subsequent discussion concerning the rationale for including each consideration at each point in time. The

rationales for the various major decision points represent a synthesis of the studies and arguments discussed in this paper.

Before making specific comments about each of the decision points, a few comments are in order about the overall design of the flowchart. One major consideration was how to minimize the impact of the ambiguity of vertical equating in the unreliable range of tests. To do this we attempted to avoid having to use the VSS as much as possible. Another major consideration was to emphasize the decision process for low scoring populations, since it is only in this population that OOL testing on a lower level would be considered.

The flowchart below has two major branches. The first branch (boxes 3 and 4) suggests that for some uses, especially program evaluation, the in-level test might be the better choice in all cases. These uses would be program evaluation that involves comparisons with the norms or other groups of students. Since for low scoring students it is difficult to tell whether the OOL or in-level derived score is more accurate, given that a test series is chosen with approximate desirable content across grades (box 1), then testing in-level avoids certain other internal validity problems in evaluation designs involving comparisons with other groups.

The second major branch of the flowchart (beginning at box 5) deals with the situation where the test scores are to be used for some other purpose than to make comparisons between groups for program evaluation. These uses involve specific test content-such as diagnosing the specific learning needs of students, or program evaluation involving measuring the progress of a group of students toward some specific skill level. In these cases the predominant concern is finding the level with the proper

content and difficulty. In designing the flowchart we had to decide whether content and difficulty could be separated--that is, content might never be judged as being appropriate unless it was of the proper difficulty. For example, a judgment of appropriate addition content would include not only whether addition of whole numbers was present but also whether the items represented adding single digits, two digits, involved carrying, etc. In the flowchart we decided to separate content and difficulty decisions. We assume that content decisions relate to finding the proper skill continuum, while difficulty decisions relate to pinpointing student ability along the skill continuum.

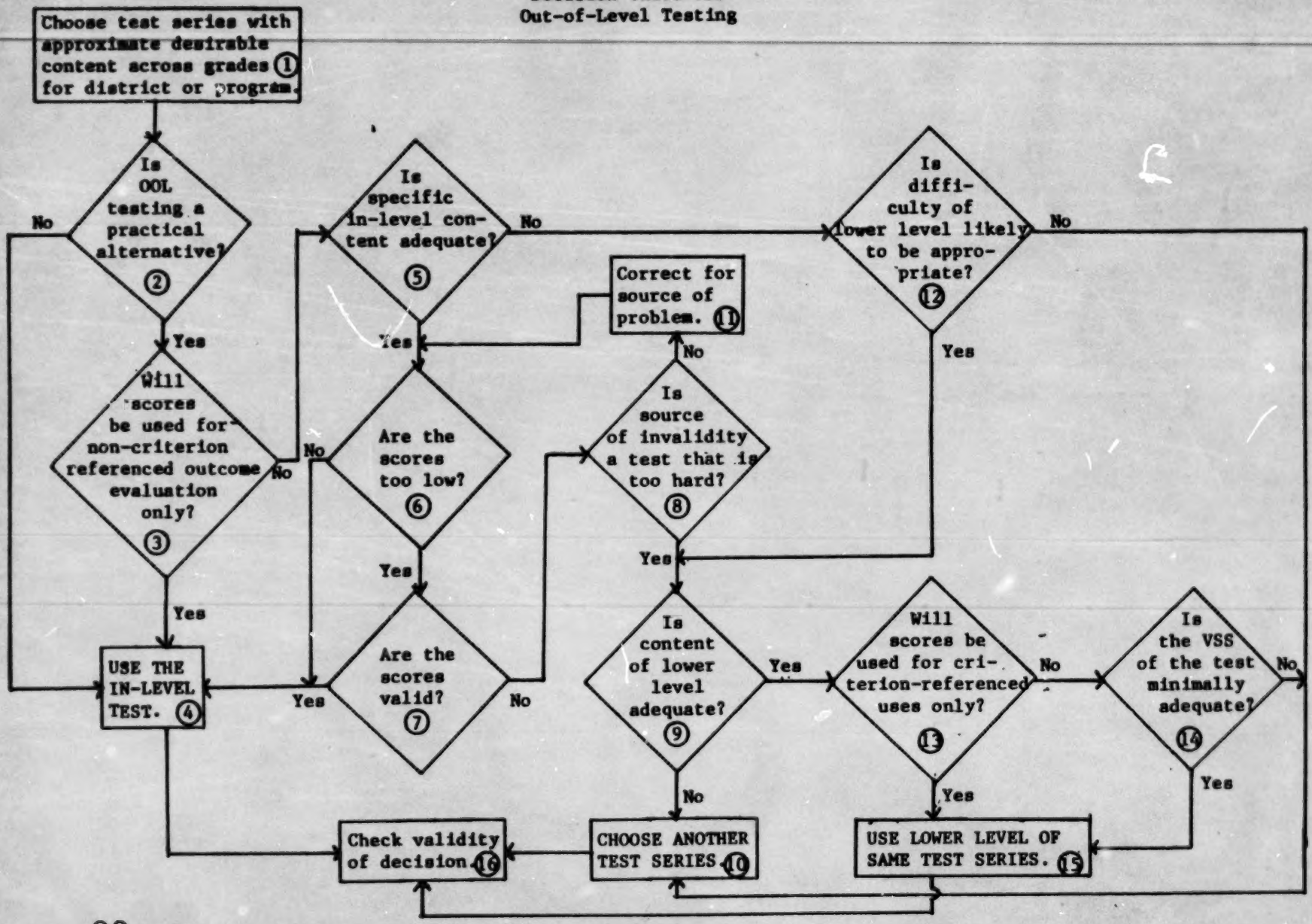
Within the second major branch are two options which become apparent in box 13. If test scores are to be used only for diagnosis and measuring progress toward content goals, then one never has to consider the adequacy of vertical equating because it is never invoked. If however, one wants to use test scores for both diagnosis and program evaluation involving comparisons between groups (especially norm referenced comparisons) then VSS adequacy needs to be considered before deciding to test OCL.

Box 1: Choose Test Series with Appropriate Content

There are several important details hidden in this step. First, in order to even consider what level of a test is most appropriate, one must be using a test series which has several levels that have been equated. Second, since it would be very difficult to choose a test series which has the specific content needed for all special programs and concerns in a district, the test series needs to have a general overall appropriateness for the district or program. Programs that have specific content needs might have to choose another testing option than a test

Arter, J.A. OOL v. In-Level Testing. AERA, 1982

Decision Chart for Out-of-Level Testing


Choose test series with approximate desirable content across grades for district or program. 1
Is OOL testing a practical alternative? 2
No
Yes
Will scores be used for non-criterion referenced outcome evaluation only? 3
No
Yes
USE THE IN-LEVEL TEST. 4
Is specific in-level content adequate? 5
No
Yes
Are the scores too low? 6
No
Yes
Are the scores valid? 7
Yes
No
Correct for source of problem. 11
Is source of invalidity a test that is too hard? 8
No
Yes
Is content of lower level adequate? 9
Yes
No
Is difficulty of lower level likely to be appropriate? 12
No
Yes
Will scores be used for criterion-referenced uses only? 13
No
Yes
Is the VSS of the test minimally adequate? 14
No
Yes
USE LOWER LEVEL OF SAME TEST SERIES. 15
CHOOSE ANOTHER TEST SERIES. 10
Check validity of decision. 16

series. Finally, the rest of the flow chart assumes that the test series chosen has a minimum amount of content relevancy. This is especially important for the sequence represented by boxes 3 and 4. (The need for content relevancy has been demonstrated by numerous studies. See for example, Jenkins & Pany, 1979; Porter, et. al., 1978; and Stevens & Rosenshine, 1977.)

Box 2: Is OOL Testing a Practical Alternative?

In many districts it would be difficult to institute an OOL testing plan. If this is the case then the more technical considerations do not matter. Things that could sidetrack an OOL testing plan would be if the state or district does not allow OOL testing, if it would be very difficult to schedule some students in a class taking one level of a test while other students were taking another level, whether there is money to purchase other test levels, and other political and emotional considerations. If OOL testing would be very difficult, then the attempt to force it through might produce more error in the data than if everyone took an in-level test. Since practicalities are very easy to ascertain, this decision point is put first. (See Gabriel, Arter and Demaline, 1980 for more complete discussion of practicalities.)

Box 3: Will Test Scores be Used for Program Evaluation Using Comparison Groups Only?

Previous opinions have been asserted that for any program evaluation effort, the content of the test to be used must match the content of the curriculum to be taught. (See, for example, Leinhardt & Seewald, 1981.) The closer the match, the better. This would seem to argue that specific content of the test in relationship to the specific materials to be taught (box 5) should be identical regardless of whether the test will be used for diagnosis or evaluation. We would like to contend that there

could be some circumstances under which problems associated with less program specific content are outweighed by other design problems. Specifically, we argue that when the sole use of a set of test scores is program evaluation involving comparison groups, the most appropriate course of action might be to always test in-level. We base this conclusion on three sets of arguments--internal validity requirements for evaluation designs using comparison groups, studies which show that low test scores are predictive of achievement, and studies which show that OOL and in-level testing cannot be relied upon to show the same derived score averages for low scoring students.

Program evaluation using comparison groups. Although there are other evaluation designs besides the three official "models" for the ESEA Title I Evaluation and Reporting System (TIERS), we will use them as examples of why internal validity considerations might always necessitate the use of in-level tests when a project is concerned with choosing the proper level from a previously identified series.

In Model A1 (the "norm-referenced model" which is used by most Title I projects) the norms of the test serve as pseudo-comparison group for the Title I students. That is, it is assumed that students will maintain their percentile status over time if they do not have special program assistance. Any gain in percentile status for participating students is assumed to be the academic impact of the program. In essence then, one is comparing the low Title I students to comparably low students in the norms, and raw test scores have to be converted to in-level percentiles regardless of the level of the test taken. If the Title I students are given an OOL test, then in essence the treatment and comparison groups are being tested with different instruments. The low

students in the norms group were also just as likely to have content and difficulty mismatches as the Title I group. The attempt to eliminate this in one group and not in the other would appear to violate good evaluation design. In addition, since the low students in the norm group probably had just as much content and difficulty mismatch, then the norms take this into account--one is really comparing two groups on a test which equally matches their needs.

Model B (the comparison group model used by Sew Title I projects) is the usual comparison group design in which the progress of Title I students is compared to that of comparable students not receiving Title I. It is assumed that the students in the two groups are the same in all educationally relevant ways. Further, it is assumed that they are receiving the same curricular content except that Title I students receive the extra Title I program. In this case it would also seem that the appropriate choice for a testing instrument would be that which has already been chosen as being appropriate for the general curriculum. If both groups are given the OOL test because its content matches the content of the Title I program better, then there is an implied mismatch to the regular curriculum that the control group is receiving. This would bias evaluation results in favor of the Title I group. If both groups are given the OOL test because the in-level test is too difficult for both groups, even though both levels have the appropriate content, then OOL testing may be more justified. One would not, of course, ever give the comparison group one level of the test and the treatment group another level.

Model C (the special regression model used by only a handful of large Title I programs) uses the group of students not selected for the Title I group as the comparison group. This means that the comparison group consists of students who score higher on the pretest than the Title I students. In this case the in-level test is the most appropriate for the comparison group in both content and difficulty. Therefore, the Title I students should also receive the in-level test.

Other considerations. The above arguments become even more convincing when the studies on the validity of chance level scores is condsidered. This literature demonstrates that even very low test scores, especially group averages, tend to be predictive measures of relative standing. We therefore feel more confident that low test scores can be compared to equally low or higher scores of a comparison group for purposes of program evaluation. The above recommendations also avoid the need to use VSSs for low scoring groups.

A final consideration for the relative placement of box 3 in relation to box 5, is that it is relatively easy to specify the uses to be made of test scores. It is harder to do a complete content and difficulty analysis of the relative test levels. Therefore, the specification of as many situations in which content and difficulty do not have to be considered is an aid to those who have to decide on test levels.

Boxes 5 and 9: Is Specific Content Appropriate?

OOL testing does not eliminate all sources of test invalidity. From a theoretical standpoint, OOL testing could be useful for eliminating test invalidity that arises from test difficulty and test content. Therefore, when test scores are to be used for diagnosis of student needs and other content relevant activities, it is crucial that the content and

difficulty merits of the various levels be considered. This is the general rationale for Boxes 8 and 12 in addition to Boxes 5 and 9.

In Boxes 5 and 9 we refer to specific content relevance, while in Box 1 we talked about general content relevance. If test scores are to be used to diagnose the instructional needs of students, or if they are to be used to measure progress toward skill objectives, then the test content needs to have a very close match to what will be taught.

The method for deciding on the match between content of test and curriculum generally consists of listing all program goals and content and then assigning each test item to each goal. Of course, some items will not match any goal. There are various rules of thumb for deciding when the overlap of items and goals is sufficient. Since these topics have been discussed in detail elsewhere (see, for example Gabriel, Arter & Demaline, 1980), we will not cover them here.

Boxes 6, 7, 8 and 11: Difficulty Considerations

If the in-level content is OK, this series of boxes looks at whether the in-level test yields test scores that are valid, even if low. Box 6 (are the test scores too low?) is a screening step. One could use the general rule of thumb flags discussed in the section on determining test score validity to screen out those suspicious scores, or group averages, that might merit close examination using a more detailed method. The closer examination of suspect scores is carried out in Box 7. If the test scores are not too low, or if the scores are low yet valid (Box 7), then one can use the in-level test (since appropriateness of content was already determined in Box 5).

To check the minimum adequacy of the VSS we would suggest two sources. The first would be for tests that have out-of-level norms (that is, a test which is considered to be OOL for a group has been given to that group anyway so that empirical norms could be developed). In this case, one could plot VSS against percentiles for the two sets of empirical norms--a percentile derived from the out-of-level empirical norms, versus a percentile obtained by converting the out-of-level score to a VSS and then to a percentile using the in-level norms (Stewart & Tallmadge, 1980). VSS should correspond to the same percentiles regardless of the manner in which they are obtained.

A second would be to look at the comparability of scores between levels for students taking more than one level of the test. One should concentrate on comparability in the reliable range of both levels (i.e., both levels are of appropriate difficulty for the students). Existing studies could be used. For most major achievement tests, this equating appears to be more or less accurate. One could also double test local students to see how the test functions with a local population.

If checks such as these turn out OK, then we are willing to overlook VSS problems in the extremes of the distribution--OOL content and difficulty appropriateness are felt to outweigh VSS problems.

If in-level score derivitives of OOL test scores are to be used, then a few other cautions need to be mentioned (Ozenne, 1978; Ozenne, 1979). First, the test level should be selected carefully enough that test levels do not have to be changed between pre and posttest if at all possible. There is ample evidence that testing a group on two levels of a test yield different derived scores. Second, when levels must be changed (for example when testing occurs spring to spring), one should

If scores or averages are not valid, then before considering whether OOL testing is worthwhile, one must decide whether the cause of the invalid scores is a test that is too hard (Box 8). As we have seen in previous sections of this paper, things other than a test that is too difficult can result in invalid low test scores--test administration, illness, etc. In order to determine the source of test invalidity, one needs to invoke his/her knowledge of the students and the testing situation. For example, teachers could be queried about what level of the test they thought was of appropriate difficulty (Fuentes & Jeffress, 1978; Yoshida, 1976). If the source of the test invalidity is not a test that is too hard, then the problems in procedure need to be corrected (Box 11), and the in-level test tried again to see if test scores are still too low (back to Box 6).

If the source of the invalidity does appear to be the difficulty of the test, then the OOL test should be considered (Box 9).

Box 13: Are the Scores to be Used for Criterion-Referenced Purposes Only?

Once it is determined that the lower level of the test is more appropriate in both content and difficulty (Box 9), then we once again have to consider the way that scores will be used. If the scores will only be used for diagnosis and other criterion-referenced needs, then a step in the decision process can be avoided--one does not need to consider the adequacy of the VSS for the test (because it will not be used), and can proceed directly to Box 15.

If the test scores are to be used for program evaluation involving comparison groups as well as for criterion-referenced purposes, then the adequacy of the VSS needs to be considered before the final decision can be made to test OOL. If the VSS is not minimally adequate, then another test or test series would be most appropriate (Box 10).

OOL testing should occur when OOL testing is a practical alternative, and:

1. The specific test content is important for diagnostic or other reasons, and

   The specific in-level content is not adequate, and

   The difficulty of the lower level is likely to be appropriate, and

   The content of the lower level is adequate.

2. The specific test content is important for diagnostic or other reasons, and

   The scores on the in-level test are not valid becuase the test is too difficult, and

   The content of the lower level of the test is adequate.

3. Test scores will be used for diagnostic purposes as well s for program evaluation involving groups, and

   The specific in-level content is not adequate, and

   The difficulty of the lower level is likely to be appropriate, and

   The content of the lower level is adequate, and

   The vertical equting is adequate.

4. Test scores will be used for diagnostic purposes as well as for progrm evaluation involving groups, and

   The scores on the in-level test are not valid because the test is too difficult, and

   The content of the lower level of the test is adequate, and

   The vertical equting is adequate.

Another test or test series should be chosen when:

1. The specific test content is important for dignostic or other reasons, and

   The content of neither the in-level nor the OOL test is adequate.

2. The specific test content is important for diagnostic or other reasons, and

   The test level with the appropriate content does not have the appropriate difficulty.

try to determine the relationship between scores on the two levels for their own students. This might involve double-testing a small group of students on both levels of the test at posttest time. Third, one should try to avoid combining derived scores for students taking different levels of a test within a grade and project (Wilson & Donlon, 1980). The single most consistent result in the studies on FLT is that one can never tell how comparable scores are between levels.

Box 16: Check Validity of Decision

Regardless of whether one finally decides to use an in-level or OOL test in a test series, or another test series entirely, one should always go back and examine the results of that decision by looking at the distributions of test scores that are obtained.

Summary

The flow chart can be summarized by listing our best recommendations for the circumstances under which in-level and OOL testing might be most appropriate:

In-level testing should occur when:

1. OOL testing is not allowed or is not practical.
2. The test series has been selected to have a content coverage that is generally appropriate across grades, and

   The only use to which scores will be put is program evaluation involving group averages and a comparison group.
3. The specific test content is important for diagnostic or other reasons, and

   The in-level test content is adequate, and

   The in-level test scores are valid, even if low.

3. Test scores will be used for diagnostic purposes as well as for program evaluation involving groups and

The test series either does not hve a VSS or the vertical equating is inadequate.

REFERENCES

Angoff, W. H. The validity of negative scores in the Scholastic Aptitude Test. Unpublished memorandum, 1964. (Cited in Boldt, R. F., Educational and Psychological Measurement, 1968, 28, 47-60.)

Applebaum, W. R. Effects of off-level assignment on ITBS grade equivalent scores. Paper presented at AERA, San Francisco, 1979.

Armbruster, B.B., Stevens, R.J., & Rosenshine, B. Analyzing content coverage and emphasis: A study of three curricula and two tests. Center for the Study of Reading, University of Illinois at Urbana, Champaign, Technical Report No. 26, 1977.

Ayrer, J. E., & McNamara, T. C. Survey testing on an out-of-level basis. Journal of Educational Measurement, 1973, 10, 79-84.

Barnes, J. A report on individualized testing in Atlanta. In Putting current ideas in measurement to work in the schools, proceedings from the Thirteenth Annual Houghton Mifflin Measurement Conference, Iowa City, Iowa, 1977.

Bianchini, J.C. & Loret, P.G. Anchor test study. Final report. Project report and volumes 1 through 30, and Anchor Test supplement, final report volumes 31 through 33, 1974. (ERIC Documents Reproduction Service Numbers ED 092601 through ED 092634).

Boldt, R. F. Study of linearity and homoscedasticity of test scores in the chance range. Educational and Psychological Measurement, 1968, 28, 47-60.

Butera, B. J., & Raffeld, P. C. Investigation of vertical scaling for out-of-level testing. Paper presented at AERA, San Francisco, 1979.

Cliff, R. The predictive value of chance-level scores. Educational and Psychological Measurement, 1958, 28, 607-613.

Crowder, C. R. Relation of out-of-level testing to ceiling and floor effects on third and fifth grade students. Paper presented at AERA, Toronto, 1978.

Divgi, D.R. Evaluation of scales for multilevel test batteries. Paper presented at AERA, Boston, 1980.

Doherty, W. Restandardization study. System Development Corporation, Santa Monica, California, 1976.

Donlon, T. F. Whose zoo? Fry's orangoutang score revisited. The Reading Teacher, October, 1971.

Donlon, T. F., & Fischer, F. E. An index of an individual's agreement with group-determined item difficulties. Educational and Psychological Measurement, 1968, 28, 105-113.

Donlon, T.F., & Rindler, S. Consistency of item difficulty for individuals and groups in the Graduate Record Examination. Paper presented at AERA, San Francisco, 1979.

Durost, W. N. Comment by. The Reading Teacher, 1971, 24, 362.

Ebel, R.L. Blind guessing on objective achievement tests. Journal of Educational Measurement, 1968, 5, 321-325.

Farr, F. Reading: What can be measured? ERIC/CRIER Reading Review Series, International Reading Association Research Fund, Newark, Delaware, 1969.

Fischer, F. E. Some properties of the personal biserial index. Journal of Educational Measurement, 1970, 7, 275-277.

Fisher, J. A. The use of out-of-grade tests with retarded and accelerated readers. Doctoral dissertation, State University of Iowa, 1961, Univ. Microfilms 61-5564.

Fry, E. The Orangoutang score. The Reading Teacher, 1971, 24, 360-361.

Fuentes, E.J., & Jeffress, B.L. The use of teacher estimates of children's achievement in selecting appropriate test levels vs. blanket out-of-level testing: The problem of an efficient method of matching the student's functional level with the proper test level. Paper presented at AERA, Toronto, 1978.

Gabriel, R., Arter, J., & Damaline, R. Functional level testing. Workshop package, Title I Technical Assistance Centers, Northwest Regional Educational Laboratory, Portland, OR 1980.

Haenn, J. F., & Proctor, D. C. A practitioner's guide to out-of-level testing. AERA, Toronto, 1978.

Hills, J.R., & Gladney, M.B. Predicting grades from below chance test scores. Journal of Educational Measurement, 1968, 5, 45-53.

Jenkins, J.R., & Pany, D. Standardized achievement tests: How useful for special education? Exceptional Children, 1978 44, 448-453.

Leinhardt, G., & Seewald, A.M. Overlap: What's tested, what's taught? Journal of Educational Measurement, 1981, 18, 85-96.

Levine, R., & Lord, F. M. An index of the discriminating power of a test at different parts of the score range. Educational and Psychological Measurement, 1959, 19, 497-503.

Long, J. V. Out-of-level survey testing in Title I--Further empirical examination. Paper presented at AERA, San Francisco, 1979.

Long, J.V., Schaffran, J. A., & Kellogg, T. M. Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. Journal of Educational Measurement, 1977, 14, 203-213.

MacGinitie, W. H. Comment by. The Reading Teacher, 1971, 24, 363.

MacRae, D. J., & Green, D. R. Comment by. The Reading Teacher, 1971, 24, 364.

McCormick, B. Draft of Out-of-level study. California State Department of Education, OPER, 1979.

Mendro, R. A report on individualized testing in Dallas. In Putting current ideas in measurement to work in the schools. Proceedings from the Thirteenth Annual Houghton Mifflin Measurement Conference, Iowa City, Iowa, 1977.

Ozenne, D. The effect of vertical scaling imprecision in the estimation of Title I project gains. Paper presented at AERA, Toronto, 1978.

Ozenne, D. You may not get what you think you are getting: What test scores mean in out-of-level testing. Paper presented at AERA, San Francisco, 1979.

Pelavin, S. H., & Barker, P. A study of the generalizability of the results of a standardized achievement test. Paper presented at AERA, San Francisco, 1976.

Porter, A.C., Schmidt, W.H., Floden, R.E., & Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-539.

Powers, S. Implications of out-of-level testing for ESEA Title I students. Paper presented at AERA, Toronto, 1978.

Raffeld, P. Functional level testing and the content shift. Paper presented at AERA, San Francisco, 1979.

Roberts, O. H. Out-of-level testing. ESEA Title I Evaluation and Reporting System, Technical Paper No. 6, RMC Research Corp., Mountain View, California, 1976.

Roberts, O.H. Out-of-level testing. In Evaluator's References, RMC Research Corp., Mountain View, California, 1980.

Schmidt, J., Liberty, P., & Butera, B. Four methods of estimating the rate of guessing on standardized achievement tests. Paper presented at AERA, San Francisco, 1979.

Slaughter, H. B., & Gallas, E. J. Will out-of-level norm-referenced testing improve the selection of program participants and the diagnosis of reading comprehension in ESEA Title I programs? Paper presented at AERA, Toronto, 1978.

Slinde, J. A., & Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Stammon, D. Chance score identification. Region VI TAC, Powell Associates, Inc., Austin, TX, 1979 Draft.

Stewart, B.L. Identifying potentially invalid scores in Title I evaluation. RMC Research Corp., Mountain View, California, 1981.

Stewart, B. L., & Tallmadge, G. K. Interlevel articulation and Title I evaluation. Paper presented at AERA, Boston, 1980.

Stewart, B.L. Discussion: Assessing the reliability and validity of chance-level scores. New Directions for Testing & Measurement, 1980b, 8, 55-58.

Tallmadge, G.K., & Horst, D.P. The use of different achievement tests in the ESEA Title I evaluation system. Paper presented at AERA, Toronto, 1978.

Temp, G. E. Out-of-level testing works. Paper presented at AERA, San Francisco, 1979.

Test selection. Workshop package, Title I Evaluation Technical Assistance Centers, Northwest Regional Educational Laboratory, Portland, OR, 1979.

Wick, J. W. A report on individualized testing in Chicago. In Putting current ideas in measurement to work in the schools, proceedings from the Thirteenth Annual Houghton Mifflin Measurement Conference, Iowa City, Iowa, 1977.

Wick, J. W., & Ward, F. Testing students at functioning reading level: A two-year report from Chicago. Unpublished paper, Board of Education, City of Chicago, Department of Research and Evaluation, 2021 N. Burling Street, Chicago, Illinois, 60614, February 14, 1977.

Wilson, K.M., & Donlon, T.F. Toward functional criteria for functional-level testing in Title I evaluation. New Directions for Testing & Measurement, 1980, 8, 51-54.

Yoshida, R.K. Out-of-level testing of special education students with a standardized achievement battery. Journal of Educational Measurement, 13, 1976, 215-221.

Young, L. R. An investigation of the scaling of the Metropolitan Achievement Tests. Doctoral Dissertation, Fordham University, 1972. University Microfilm 73-1521.